NUI Galway
OÉ Gaillimh

Semester 1 Examinations 2021-2022

Course Instance Code(s)	4BCT1, 4BDS1, 4BMS2, 4BS2, 1MECE1, 1MEME1, SPE
Exam(s)	Fourth BSc in Computer Science & Information Technology Fourth Bachelor of Arts with Data Science Fourth Bachelor of Science (Mathematical Science) Fourth Bachelor of Science (Hons.) ME in Electronic & Computer Engineering ME in Mechanical Engineering Structured PhD
Module Code(s)	CT4101
Module(s)	Machine Learning
Paper No.	1

External Examiner(s)	Dr. Ramona Trestian
Internal Examiner(s)	Prof. Michael Madden *Dr. Patrick Mannion

Instructions:	Answer any 3 questions out of 4 questions. All questions carry equal marks (25 marks each). The total (out of 75 marks) will be converted to a percentage after marking.
Duration	2 hours
No. of Pages	5
Discipline(s)	School of Computer Science
Course Co-ordinator(s)	Dr. Colm O'Riordan (BCT), Dr. Nick Tosh (BDS), Prof. Dane Flannery (BMS), Dr. Emma Holahan (BS), Prof. Martin Glavin (MECE), Dr. Noel Harrison (MEME)

Requirements:

Release in Exam Venue	Yes [X]	No [ ]
MCQ Answersheet	Yes [ ]	No [X]
Handout	None	
Statistical/ Log Tables	None	
Cambridge Tables	None	
Graph Paper	None	
Log Graph Paper	None	
Other Materials	None	
Graphic material in colour	Yes [ ] No [X]	

PTO

Question 1 (25 marks)

Part (a)

Explain the differences between the following machine learning tasks: **classification**, **regression, clustering**. For each task type, list a specific application, and an algorithm which may be used to learn a suitable model.

[6]

Part (b)

Explain what is meant by the term **curse of dimensionality** and briefly discuss a method to reduce its effect.

[3]

Part (c)

Explain what is meant by the terms **overfitting** and **underfitting**. As part of your answer include sketches demonstrating how overfitting and underfitting can be identified in both classification and regression tasks.

[8]

Part (d)

In the context of regression tasks, provide an example of a domain independent method to measure error, and an example of a method that measures error in the same units as the dependent variable. As part of your answer explain how both of your chosen error measurement methods may be calculated (including formulae), and discuss whether there are any situations where one method is more useful than the other.

[8]

Question 2 (25 marks)

Part (a)

ID	Target	Prediction
1	TRUE	TRUE
2	FALSE	TRUE
3	FALSE	FALSE
4	TRUE	TRUE
5	FALSE	FALSE
6	FALSE	TRUE
7	TRUE	TRUE
8	TRUE	TRUE

The table above presents the results of evaluating a classifier on a test set for a binary classification task.

(i) Present the results of the evaluation above in a confusion matrix [3]
(ii) Calculate the misclassification rate of the classifier [2]
(iii) Calculate the true negative rate of the classifier [2]
(iv) Calculate the precision of the classifier [2]
(v) Calculate the recall of the classifier [2]

[11]

Part (b)

Explain what is meant by the term **cross validation**. Briefly discuss how you would use cross validation to evaluate the performance of a machine learning algorithm applied to a regression task.

[4]

Part (c)

Explain how receiver operating characteristic (ROC) curves may be used to compare the performance of different classifiers for a binary classification task. Sketch an example of an ROC curve for a binary classification task as part of your answer, clearly labelling the axes and clearly indicating a point that represents ideal performance.

[6]

Part (d)

Explain what is meant by the term **hyperparameter** in the context of machine learning algorithms. Briefly discuss how a grid search could be used to determine suitable hyperparameter values for a machine learning algorithm.

[4]

PTO

Question 3 (25 marks)

Part (a)

Describe the key similarities and key differences between the ***k*-means** algorithm and the ***k*-nearest neighbours** algorithm. As part of your answer provide a brief description of both algorithms.

[10]

Part (b)

Describe with the aid of a diagram how the elbow method may be used to identify a suitable value for *k* when applying the *k*-means algorithm to a dataset.

[4]

Part (c)

Provide an example of a distance metric that is suitable for computing similarity between instances in datasets with **discrete attributes**, and an example of a distance metric that is suitable for computing similarity between instances in datasets with **continuous attributes**. Describe how each of your chosen distance metrics may be calculated with the aid of simple original examples.

[6]

Part (d)

Explain with the aid of a simple original example dataset (i.e., one not covered in the lecture notes) why it is often necessary to use normalisation with the k-nearest neighbours algorithm. Describe how z-normalisation may be applied to your example dataset.

[5]

Question 4 (25 marks)

Part (a)

Explain the differences between the following types of activation functions in the context of neural networks:

(i) Threshold
(ii) Logistic
(iii) ReLu

[6]

Part (b)

Explain what is meant by the term **one-hot encoding**. Include a simple example as part of your explanation.

[4]

Part (c)

Explain the following concepts in your own words:

(i) Entropy [2]
(ii) Information gain [2]

[4]

Part (d)

Describe in detail (including pseudocode) an algorithm that could be used to learn a decision tree for a classification task.

[8]

Part (e)

Briefly describe a method that could be used to combat overfitting in a decision tree.

[3]

END OF EXAM PAPER